WANDERER DOCUMENTATION

1. AIM

Wanderer is a very simple and intuitive web tool allowing real time access and visualization of gene expression and DNA methylation profiles obtained from the TCGA Research Network (http://cancergenome.nih.gov/) using gene targeted queries. Wanderer is addressed to a broad variety of experimentalists and clinicians without deep bioinformatics skills.

2. ACCESS AND CITATION

Wanderer may be accessed at http://www.maplab.cat/wanderer

If you find this software useful please consider citing our paper: (in preparation).

3. USING WANDERER

3.1. Enter a gene name (BRCA1 is used as default) or the equivalent Ensembl gene ID and press the Refresh button to update the output. Modifying the rest of options will automatically update the graphs.

3.2. Choose a Dataset (default dataset is Breast Invasive Carcinoma (BRCA)) and the Data type (450k Methylation array or Illumina HiSeq RNAseq) in the respective drop down menus.

3.3. Use the Zoom sliding bar to adjust the displayed chromosomal region or tick the Specify a region box for a more precise selection. Only values in the range defined by the zoom slider will be accepted.

3.4. Customize the graphical display of the data using Plotting parameters. Graphs are automatically updated after any change.

3.5. Download graphs and tables containing the selected data or share the output by generating a permanent link.

4. DESCRIPTION OF THE DOWNLODABLE FILES GENERATED BY WANDERER

Wanderer generates several images and data files that may be downloaded as a compressed file. The files are named using the following nomenclature:

```
Wanderer_[dataset]_[data type]_[date and time of the query].
```

4.1 DNA METHYLATION FILES

The predefined query for DNA methylation produces the following graphs and data files:

4.1.1. Profile plots displaying the beta values for each probe in two separate panels representing subset of normal (upper panel, blue marks) and tumor samples (bottom panel, red marks) of the selected dataset. Lines link different probes corresponding to the same sample. The gene location and direction is depicted with an arrow. CpG island located probes are colored in green.

File name examples:

Vectorial format: `Wanderer_BRCA1_methylation_brca_Apr_29_2015_at_170440_CEST.pdf`

Raster version: `Wanderer_BRCA1_methylation_brca_Apr_29_2015_at_170440_CEST.png`

4.1.2. Profile plots displaying the average methylation for all the normal (blue) and tumor samples (red) of the selected dataset. The gene location and strand is depicted with an arrow. CpG island located probes are colored in green. The CpGs showing statistical differences between normal and tumor are highlighted with an asterisk (Wilcoxon adjusted p-value < 0.05).

File name examples:

Vectorial format: `Wanderer_BRCA1_Mean_methylation_brca_Apr_29_2015_at_170440_CEST.pdf`

Raster version: `Wanderer_BRCA1_Mean_methylation_brca_Apr_29_2015_at_170440_CEST.png`

4.1.3. Two tables consisting of a comma separated data matrix containing data for normal and tumor samples respectively. Table contents: First row: Sample ID, First column: 450K Methylation array probe name, second column: DNA methylation beta value.

File name examples:

`Wanderer_BRCA1_methylation_brca_Normal_Apr_29_2015_at_170440_CEST.csv`

`Wanderer_BRCA1_methylation_brca_Tumor_Apr_29_2015_at_170440_CEST.csv`

4.1.4. A comma separated data matrix with the annotation for each of the informative probes and descriptive analysis of the DNA methylation data. The columns correspond to:

- probe, probe name
- chr, the chromosome
- cg_start, the genomic position the CpG starts at
- cg_end, the genomic position the CpG ends at
- percentgc, the GC content of the illumina 450k array probe
- probetype, the type of the illumina 450k array probe
- probestart, the genomic position the probe starts at
- probeend, the genomic position the probe ends at
- genestart, the genomic position the closest gene to the probe starts at
- geneend, the genomic position the closest gene to the probe ends at
- genestrand, the closest gene strand
- ENSEMBL_geneID, the closest gene id at ensembl
- genebiotype, the closest gene biotype (protein coding, retained intron...)
- genename, the closest gene symbol
- cpgistart, for those CpG are inside a CpG island, the coordinate this islands starts at
- cpgiend, for those CpG are inside a CpG island, the coordinate this islands ends at
- cpgiid, for those CpG are inside a CpG island, CpG island identifier
- Norm_nsamples, number of normal samples for this dataset in this release
- Norm_mean, mean of the beta values for normals
- Norm_sd, standard deviation of the beta values for normals
- Tum_nsamples, number of tumor samples for this dataset in this release
- Tum_mean, mean of the beta values for tumors
- Tum_sd, standard deviation of the beta values for tumors
- wilcox_stat, Wilcoxon Rank Sum Test W parameter (nonparametric comparison of normals vs tumors)
- pval, Wilcoxon Rank Sum Test p value (nonparametric comparison of normals vs tumors, low values indicates that differences are detected)
- adj.pval, Benjamini and Hochberg adjustement (False Discovery Rate) for multiple testing.

File name examples:

`Wanderer_BRCA1_methylation_brca_annotations_and_statistical_analysis_Apr_29_2015_at_170440_CEST.csv`

4.2 GENE EXPRESSION FILES

4.2.1. Profile plots displaying the log2-transformed RPKM values (Guo, 2013) for each exon in two separate panels for a subset of normal (upper panel, blue marks) and tumor samples (bottom panel, red marks) from the selected dataset. Lines link values corresponding to the same sample. The gene location and strand is depicted with an arrow.

File name examples:

Vectorial format: `Wanderer_BRCA1_expression_brca_Apr_29_2015_at_170440_CEST.pdf`

Raster version: `Wanderer_BRCA1_expression_brca_Apr_29_2015_at_170440_CEST.png`

4.2.2. Profile plots of the average expression for all the normal (blue) and tumor samples (red) of the selected dataset. The gene location and strand is depicted with an arrow. The exons showing statistically significant differences between normal and tumor are highlighted with an asterisk (Wilcoxon adjusted p-value < 0.05).

File name examples:

Vectorial format: `Wanderer_BRCA1_Mean_expression_brca_Apr_29_2015_at_170440_CEST.pdf`

Raster version: `Wanderer_BRCA1_Mean_expression_brca_Apr_29_2015_at_170440_CEST.png`

4.2.3. Boxplot and a stripchart graphs showing the expression values summarized by gene for all the normal (blue) and tumor samples (red) of the selected dataset. The expression values are log2-transformed normalized RSEM values (Guo, 2013) and reflect the expression of the gene as a whole.

File name examples:

Vectorial format: `Wanderer_BRCA1_boxplot_expression_brca_Apr_29_2015_at_170440_CEST.pdf`

Raster version: `Wanderer_BRCA1_boxplot_expression_brca_Apr_29_2015_at_170440_CEST.png`

4.2.4. Two tables consisting of a comma separated data matrix with the exon location (first column) and a column with the log2-transformed RPKM values for each of the exons of all the samples in the dataset. Sample names are in the first row (header).

File name examples:

`Wanderer_BRCA1_expression_brca_Normal_Apr_29_2015_at_170440_CEST.csv`

`Wanderer_BRCA1_expression_brca_Tumor_Apr_29_2015_at_170440_CEST.csv`

4.2.5. A comma separated data matrix with the available normal and tumors sample names (first column) and a column with the log2-transformed RSEM values for your gene of interest.

File name examples:

`Wanderer_BRCA1_expression_brca_Normal_RNAseqGENE_Apr_29_2015_at_170440_CEST.csv`

`Wanderer_BRCA1_expression_brca_Tumor_RNAseqGENE_Apr_29_2015_at_170440_CEST.csv`

4.2.6. A comma separated data matrix with the annotation of your gene expression data, as well some descriptive analysis. The columns correspond to:

- exon, exon identifier according to the TCGA pipeline
- id, exon identifier according to Genome Browser exons track
- ENSEMBL_geneID, the gene identifier according to ENSEMBL (ENSG identifier)
- ENSEMBL_transcriptID, the transcript identifier according to ENSEMBL (ENST identifier)
- chr, the chromosome the exon is located at
- exon_start, the genomic coordinate the exon starts at
- exon_end, the genomic coordinate the exon ends at
- strand, the genomic strand of the exon's gene.
- genestart, the genomic start position of the exon's gene.
- geneend, the genomic end position of the exon's gene.

- genebiotype, exon's gene biotype (protein coding, retained intron...)
- genename, exon's gene symbol
- rnaseqgeneid, the TCGA exon's gene identifier
- Norm_nsamples, number of normal samples for this dataset in this release
- Norm_mean, mean of log2-transformed RPKM for normals
- Norm_sd, standard deviation of log2-transformed RPKM in normals
- Tum_nsamples, number of tumor samples for this dataset in this release
- Tum_mean, mean of log2-transformed RPKM for tumors
- Tum_sd, standard deviation of log2-transformed RPKM in tumors
- wilcox_stat, Wilcoxon Rank Sum Test W parameter (nonparametric comparison of normals vs tumors)
- pval, Wilcoxon Rank Sum Test p value (nonparametric comparison of normals vs tumors, low values indicates that differences are detected)
- adj.pval, Benjamini and Hochberg adjustement (False Discovery Rate) for multiple testing.

File name example:

`Wanderer_BRCA1_expression_brca_annotations_and_statistical_analysis_Apr_29_2015_at_170440_CEST.csv`

4.3 CLINICAL DATA

TCGA offers the clinical data as biotab files (among other formats). Biotab files are an amenable, spreedsheet-friendly data sources that reflect clinical and biospecimen information for a set of patients. In this file, each column is a clinical element and each row, a TCGA participant.

We note that biotab files cover data unique to each TCGA participant (for instance, the patient's weight and height) and therefore are independent on whether methylation or expression was scrutinized. We offer the biotab files for all the participants available for a dataset (for instance, colon adenocarcinoma), regardless of whether they were explored for expression or methylation, nor in solid primary tumors or in adjacent normal, and therefore the final user is asked to filter the TCGA participants according to his/her needs.

File name example:

`BRCA_Clinical__nationwidechildrens.org_clinical_patient_brca.txt`

5. DATA AVAILABILITY

Wanderer uses a TCGA snapshot taken on July 2014. Data availability for normals (N) and tumors (T) is detailed below.

Wanderer offers external links to Genome Browser, Regulome Explorer and cBioPortal. We note that some of the user's queries might involve a gene or a dataset (i.e. mesothelioma) that was not covered in either Regulome Explorer or cBioPortal. We note that the TCGA participants used in the analysis might differ between Wanderer and these others Web resources, as they used different data snapshots from TCGA. Moreover, Regulome Explorer datasets might contain tumor-only data (without the normal adjacent).

Description	Name	Meth_N	Meth_T	RNAseq_N	RNAseq_T	Regulome Explorer	cBioP
Adrenocortical carcinoma	ACC	0	80	0	79	acc_2015_03_31	acc_t
Bladder Urothelial Carcinoma	BLCA	21	358	19	267	blca_20may13_test	blca_
Brain Lower Grade Glioma	LGG	0	511	0	513	lgg_04oct13_seq	lgg_t
Breast invasive carcinoma	BRCA	98	743	113	1052	brca_manuscript_rerun_nov12d_pw	brca_
Cervical squamous cell carcinoma and endocervical	CESC	3	256	3	207		cesc_

Description adenocarcinoma	Name	Meth_N	Meth_T	RNAseq_N	RNAseq_T	Regulome Explorer	cBioP
Colon adenocarcinoma	COAD	38	302	41	192	coad_03feb13_seq_tumor_only	coadr
Esophageal carcinoma	ESCA	16	185	0	0		esca_
Glioblastoma multiforme	GBM	2	129	5	156	gbm_2013_pub_tumor_only	gbm_
Head and Neck squamous cell carcinoma	HNSC	50	528	43	497	hnsc_03feb13_seq_tumor_only	hnsc_
Kidney Chromophobe	KICH	0	66	25	66		kich_
Kidney renal clear cell carcinoma	KIRC	160	324	72	518	kirc_01oct12_A_pw	kirc_t
Kidney renal papillary cell carcinoma	KIRP	45	226	30	198		kirp_
Liver hepatocellular carcinoma	LIHC	50	256	50	212		lihc_t
Lung adenocarcinoma	LUAD	32	463	58	488	luad_03feb13_seq_tumor_only	luad_
Lung squamous cell carcinoma	LUSC	43	361	50	491	lusc_03feb13_seq_tumor_only	lusc_
Lymphoid Neoplasm Diffuse Large B+AC0-cell Lymphoma	DLBC	0	48	0	28		dlbc_
Mesothelioma	MESO	0	37	0	36		
Ovarian serous cystadenocarcinoma	OV	0	10	0	262	ov_03feb13_ary_tumor_only	ov_tc
Pancreatic adenocarcinoma	PAAD	10	146	4	96		paad_
Pheochromocytoma and Paraganglioma	PCPG	3	179	3	178		pcpg_
Prostate adenocarcinoma	PRAD	49	340	52	374		prad_
Rectum adenocarcinoma	READ	7	98	9	72		coadr
Sarcoma	SARC	4	242	2	103		sarc_
Skin Cutaneous Melanoma	SKCM	2	92	1	82		skcm_

Description	Name	Meth_N	Meth_T	RNAseq_N	RNAseq_T	Regulome Explorer	cBioP
Stomach adenocarcinoma	STAD	2	339	0	0	stad_23jan14_seq_tumor_only	stad_
Thyroid carcinoma	THCA	56	507	59	498	thca_18oct14_TP	thca_
Uterine Carcinosarcoma	UCS	0	57	0	57		ucs_t
Uterine Corpus Endometrial Carcinoma	UCEC	46	438	11	370	ucec_28jun13b_seq_tumor_only	ucec_
Uveal Melanoma	UVM	0	80	0	0	uvm_20150515_private	

6. FREQUENTLY ASKED QUESTIONS

6.1. Wanderer says there are not enough samples to perform statistical analysis, what does it mean?

This means there were not enough number of informative samples to compute the Wilcoxon test. We note that, although we calculate the test with small number of samples, this result might be meaningless without a minimum number of cases.

6.2. The CSV files are a single line file!

We use Linux/UNIX linefeeds. Although your text viewer (i.e. notepad) might merge all the lines under Windows, your spreadsheet software (i.e. Excel) will properly recognize the line separation.

6.3. How can I open a CSV in my spreadsheet software?

In most cases it should work by just using the open with (right click with the mouse on the file icon). Alternatively, launch the application and import data from text, comma-separated or character-delimited values (this depends on the software, i.e. LibreOffice).

6.4. How and when did you download the data from the TCGA?

Data from TCGA was downloaded using the TCGA-Assembler (Zhu, 2014) on July 8th 2014.

6.6. What is the origin of annotation data?

450k methylation array probes have been annotated as described in (Price, 2013).

Expression data annotation was performed using gene annotations and exon positions as provided by the TCGA and UCSC Genome Browser. Data was fetched using BioMart and, when needed, was merged according to feature overlaps (Diez-Villanueva, 2015). For methylation data, we took advantage of the chip annotation by (Price, 2013).

6.6. What do RPKM and RSEM stand for?

Both values are directly obtained from the TCGA pipeline for RNA seq v2 expression analysis. RPKM stands for *reads per kilobase per million mapped reads* (Guo, 2013) and is used to represent exon expression levels. RSEM stands for *RNA-Seq by Expectation-Maximization* (Guo, 2013) and is used to represent gene expression levels. We log2-transform these values adding one to get rid of zeroes; that means we plot log_2(x+1) RPKM or RSEM values.

7. REFERENCES

(Diez-Villanueva, 2015) Association analysis of genomic regions based on permutation tests

(Guo, 2013) Guo, Yan, et al. Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data. PloS one 8.8 (2013): e71462.

(Price, 2013) Price et al. Additional annotation enhances potential for biologically-relevant analysis of the

Illumina Infinium HumanMethylation450 BeadChip array. Epigenetics & Chromatin 2013, 6:4 doi:10.1186/1756-8935-6-4.

(Zhu, 2014) Y. Zhu, P. Qiu, Y. Ji. TCGA-Assembler: Open-Source Software for Retrieving and Processing TCGA Data. Nature Methods. 11:599-600, 2014. | doi:10.1038/nmeth.2956).

8. VERSIONS AND COPYRIGHT

This documentation corresponds to Wanderer v1.0. The full development log can be found at https://sourceforge.net/projects/tcga-wanderer/. Wanderer is free software and is provided under the GPL v2 terms.

The previous major releases are:

- Wanderer v1.0
- Alpha Release Mon Dec 29 15:14:48 2014 (commit 21f1c7d9c742011e289842752a1ea871c13dc5fb)
- Project start Mon Oct 6 15:04:45 2014 (commit 3e5440a552d04319331415e74635e144045cc6ae)

This document corresponds to

```
commit ec5c6643247d07eaa14251bda719d9fd4e56cfee Author: Izaskun Mallona <imallona@imppc.org> Date:
Tue May 26 12:40:54 2015 +0200
```

Some datasets have limitations for usage until a global analysis is published; please contact TCGA before publishing.

9. CONTACT

Feel free to contact Anna Diez-Villanueva (adiez@imppc.org), Izaskun Mallona (imallona@imppc.org) or Miguel A. Peinado (map@imppc.org).

We provide more tools and data at http://www.maplab.cat.